DOCUMENT RESUME

ED 110 514 TM 004 787

AUTHOR Popham, W. James
TITLE Program Fair Evaluation--Summative Appraisal of Instructional Sequences with Dissimilar Objectives.
INSTITUTION Southwest Regional Laboratory for Educational Research and Development, Los Alamitos, Calif.
PUB DATE 16 Sep 68
NOTE 9p.

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS *Classification; *Comparative Analysis; Curriculum Evaluation; *Educational Objectives; Evaluation; *Instructional Programs; Models; *Program Evaluation; Summative Evaluation

ABSTRACT

A comparative evaluation involving two instructional programs is given, although the approach can easily serve to compare more than two programs. The steps involved in conducting a program fair evaluation of two instructional programs are: (1) Identify objectives (a) common to both programs, (b) unique to one program, and (c) unique to the other program; (2) Collect or construct test items based on the three sets of objectives; (3) Combine the test items into a three-part examination; (4) Assign estimates of importance, as explicitly as possible, to the three sets of objectives; (5) Administer each of the instructional programs to one of two randomly selected groups of appropriate learners; (6) Administer the three-part examination to both groups; and (7) Appraise results and reach a decision regarding the preferred program. The summative evaluation scheme described here is relatively straightforward and can, therefore, be carried out with little difficulty. It can be applied to the evaluation of short-duration instructional sequences or to programs of much greater magnitude. While conceptually simple, it provides the evaluator with opportunities to employ sophisticated quantification schemes to deal with such problems as value weightings of objectives. The major purpose of the procedure is to make evaluators attentive to the differential consequences of employing instructional programs which were designed with different intentions. (Author/BJG)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED110514

SOUTHWEST REGIONAL LABORATORY
RESEARCH MEMORANDUM

TM

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY

SWRL

TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER.

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

DATE: 16 September 1968
RELEASE: Robert L. Baker RLB

PROGRAM FAIR EVALUATION--SUMMATIVE APPRAISAL OF INSTRUCTIONAL SEQUENCES WITH DISSIMILAR OBJECTIVES

W. James Popham

Even as today's educational evaluators acquire increasing sophistication in appraising instructional programs, they encounter classes of evaluation problems which tax their expertise. Some of these problems are associated with relatively new issues, while others have been with us for many years. One question receiving considerable attention in the 1960's pertains to comparative evaluation of two or more instructional sequences which have some objectives that are the same but some that are different.

At first glance this would seem to be an age-old evaluation problem, since educators for many years have had to choose among competing instructional sequences. Yet, only recently has it been possible to apply vigor to the task. This stems from a development in the field of instruction, namely, the tendency to prepare instructional sequences which are essentially replicable and which take responsibility for promoting specified changes in appropriate learners. In earlier days curriculum specialists usually designed what they hoped was an optimal instructional program, then evaluated it. Rarely did they have an opportunity to choose among competing instructional sequences. It was

TM004 787

Permission to reprint or quote from this working document, wholly or in part, should be obtained from SWRL, 11300 La Cienega Blvd., Inglewood, Calif., 90304.

tough enough just to get one into operation. But today, due in part to the impact of the programmed instruction movement, more and more replicable instructional products are appearing on the educational market. These products range from short, one-lesson programs to elaborate, year-long packages. The replicability of these products makes it possible for school people to select one with the assurance that it will be effective in the local situation. Local curriculum workers need not put up their own instructional preserves, they can now buy them at the market.

Comparative Evaluation

But as any housewife will tell you, comparative shopping is difficult. It's hard to decide which of two roasts to purchase when there are several differences between them, e.g., quality (prime versus choice), weight (4 lbs., 6 ozs, versus 5 lbs., 2 ozs.), bone content ("not much" versus "mostly marrow"). Decisions are particularly difficult when such differences do not uniformly favor one of the potential purchases. In the same way, a consumer of educational products faces a complex problem when he is obliged to choose among sets of instructional materials which, among other differences, are designed to promote goals that are--at least to some extent--dissimilar.

Let us consider, for example, an increasingly common kind of educational dilemma. A local school district curriculum committee has been commissioned to select a set of biology materials. There are now several sets of such materials on the market, each with its own texts, work books, and teacher manuals. Because of careful structuring of the materials, it can be said they are essentially replicable, that is, can probably be used in roughly the same way by various biology instructors. Each of the three most eligible sets of materials has been field-tested to the point where data are available on the degree to which each program promotes achievement of its objectives. However, the objectives of the three programs differ, so even with a wealth of performance data on each program, the committee must still somehow discriminate among the programs with respect to intended outcomes.

Selecting a program is a summative evaluation task.[1] Such tasks will be presented to educators with increasing frequency as a consequence of the expanding availability of replicable instructional products. How should such tasks best be handled?

Inadequate Strategies

Approaches range from the crude to the complex. An approach will be advocated here which, it is hoped, is (1) sufficiently practicable

---

[1] M. Scriven, "The Methodology of Evaluation." in B. O. Smith (Ed.) Perspectives of Curriculum Evaluation. Chicago: Rand McNally, 1967.

that it can be readily used and (2) sufficiently sensitive that it yields the appropriate information for evaluative decisions. But first let's dispense with some of the more simplistic evaluation strategies which have, unfortunately, been used all too often in comparing instructional sequences with differing objectives.

It should be noted at the outset that this discussion pertains to comparative evaluations (i.e., judging which contender is best) among instructional sequences that employ replicable methods and yield replicable results. For example, this includes sets of self-instruction programs in which a "live" teacher plays little or no role; and instructional sequences in which teachers are heavily involved but are also heavily influenced as to the form of the instruction they produce. It obviously doesn't make much sense to devote intense efforts to evaluating non-replicable instruction, such as that spontaneously generated by "off-the-cuff" teachers whose teaching is so variable that generalizations regarding their future performances are risky. So we are talking here about comparative evaluations of such entities as long-term national curriculum projects which yield fairly reliable instruction (perhaps because they are largely material-based) or short-term instructional sequences such as the teacher's use of highly prescriptive lesson plans and unit plans.

Some evaluators respond exclusively to "presentation" stimuli which may or may not be related to student attainment of instructional objectives. For instance, educators often choose one set of reading materials over another because of packaging: illustrations, covers, and page make-up. Such things are not necessarily irrelevant to learner achievements but they usually have not been demonstrated to be germane. Again, we find some educators preferring one program to another primarily because they are impressed with an author's style or with the way he treats a particular concept. Such approaches never raise the only truly important question: "What happens to the learner as a consequence of his encounter with the instructional sequence?" Of course style and treatment of content may be critically related to learner post-instructional behavior. But such relationships are rarely, if ever, verified. Instead, the evaluator trusts his intuition, and intuitively derived evaluations such as these are difficult to defend.

A more reasonable approach to cross-program evaluation involves use of learner performance measures. A common practice has been to administer either a standardized achievement test or a specially constructed achievement test to learners who have completed different instructional sequences, then compare their performances. The weakness with this approach, however, is that these tests often lack relevance to the objectives of the different instructional sequences. A standardized test, for example, is usually developed in a norm-referenced[2]

---

[2] R. Glaser. "Instructional Technology and the Measurement of Learning Outcomes: Some Questions." American Psychologist, XVIII, 1963, 519-521.

context where the primary purpose of the test is to differentiate among individuals, rather than in a criterion-referenced context where the primary purpose of the test is to assess the degree to which an individual meets a criterion. Because of procedures associated with the development and refinement of norm-referenced measures (e.g., item analysis and the quest for highly variant scores) such tests often fail to retain items which assess truly important intended outcomes of given instructional programs. Therefore, to use these tests as a principal basis for judgment may obscure the true effectiveness of the program in question.

In somewhat the same way a test specially designed to assess the merits of two or more discrete instructional programs often suffers from attempting to serve multiple masters. By trying to cover fairly the objectives of more than one program, the resulting test is often a watered-down instrument which tells the evaluator little. Even worse, such a test may inadvertently place greater emphasis on the objectives of one program, thereby favoring it when learner performance is evaluated.

Although these two approaches--use of criteria other than learner behavior and use of a single test--have been the most widely employed procedures for comparing two or more instructional programs with different objectives, there are other less common strategies. Wolf[3] recently described some of these and identified weaknesses in each.

Program Fair Evaluation

The proposed scheme for avoiding some of the previously identified weaknesses is both conceptually simple and easy to implement. For ease of exposition, a comparative evaluation involving only two instructional sequences will be considered, although the approach can easily serve to compare more than two programs. For a hypothetical case, consider the problem faced by an evaluator who must choose between Program X and Program Y which have some objectives that are the same and some that are different.

First, the evaluator should isolate the instructional objectives of each program. If the objectives are not presented by the programs' developers, it may be necessary to infer them from the programs' criterion tests or, lacking these, from the programs[4] themselves. For the

---

[3] R. Wolf, "Program Free Testing," Research Memorandum, Southwest Regional Laboratory for Educational Research and Development, Inglewood, California, May 16, 1968.

[4] W. J. Popham, "A Comparison of 'Rough and Dirty' Methods of Evaluating Self-Instruction Programs," Working Paper, UCLA Program Effectiveness Project, October, 1964.

evaluator's purposes, such objectives must be stated operationally (e.g., in terms of intended learner behavior changes). Loose, global objectives are of little utility to the evaluator.

Each program's objectives should then be grouped, as indicated in Figure 1, into those common to both programs, those unique to Program X, and those unique to Program Y.

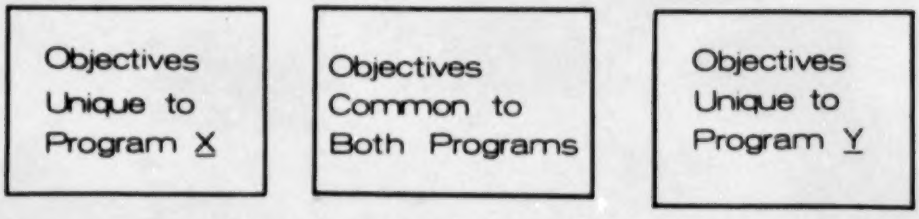

Figure 1. A classification of objectives for Program X and Program Y.

In isolating the unique and the common objectives, it may be possible to work exclusively from test items rather than objectives. Doubtless, however, it is advantageous to delineate objectives clearly since by examining them (1) the class of test items covered is more easily seen, and (2) the appropriateness of a given test item to a particular program is more easily determined.

Next, representative tests of each of the three sets of objectives must be generated. And "test" here is used broadly, in the sense that one could generate written and non-written measures for objectives in the affective domain as well as the cognitive. An ideal proportion should be sought between number of test items and number of objectives. If the situation presented in Figure 2 prevailed, for example, where one set of unique objectives was much larger than the other, about twice as many test items would be prepared for objectives unique to Program X as for objectives unique to Program Y. This is complicated, unfortunately, because measuring performance of one objective may take

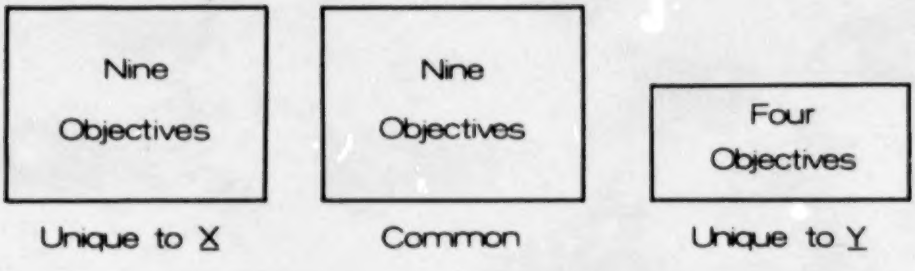

Figure 2. A hypothetical classification of program objectives.

whether common or unique. But it would be helpful at this stage even to make a gross commitment where the evaluator might assert: "Objectives unique to Program X are more important than the common objectives. Further, the common objectives are more important than those unique to Program Y."

Next, a sample of appropriate learners is randomly assigned to one of two instructional treatments and, after they have completed their respective sequences, they are given the three-part criterion test. We are assuming that the average time taken to complete each program is comparable, although adjustments could be made in criterion test scores if one program took much longer to complete than the other.

Finally, one of the two programs is selected based on the performance of the two groups of learners on the three subtests and on the previously judged worth of each set of objectives. In Figure 3, for example, it would seem that because the objectives unique to Program Y were considered so important (8 points on a 10-point scale), that Program Y would be selected here. Both programs yield learner performance on the other two sets of objectives which, though favoring Program X, are roughly comparable. Program Y produces clearly superior results on objectives unique to itself while doing almost as well as Program X on the other two dimensions.

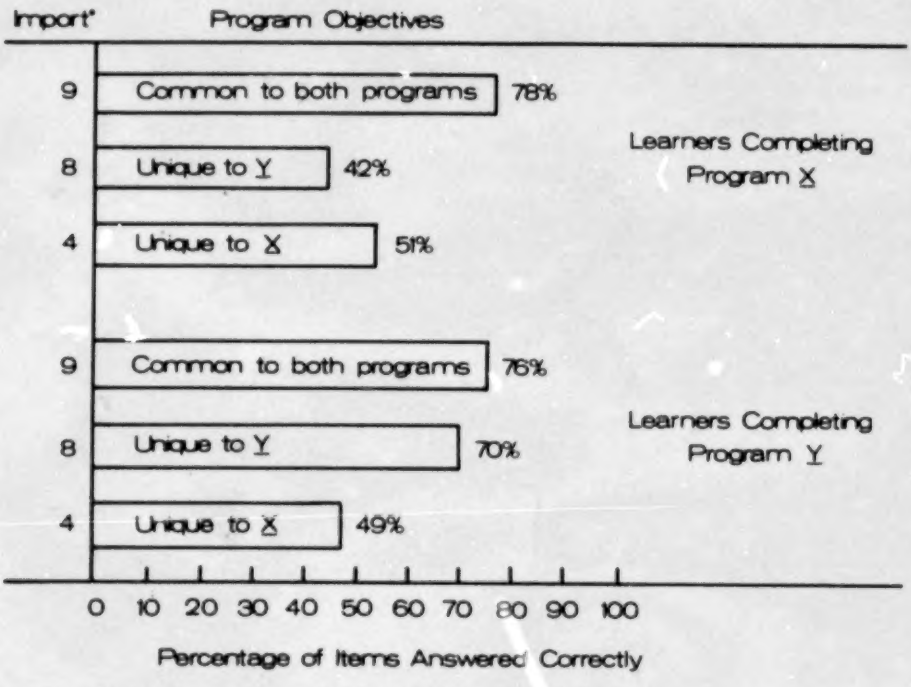

Figure 3. Hypothetical results of two groups of learners completing two different programs.

Of course, one could generate other fictitious sets of results where the decision-maker would face a more difficult choice. Considering value weightings and relative success of learners on the three sets of objectives, situations might be presented wherein the evaluator would resort to a coin flip. But even the most elaborate evaluation scheme will occasionally yield such impasses. The value of the procedure described here is that it can sharpen the evaluator's awareness of the degree to which different instructional sequences perform different kinds of jobs.

Results on such an examination must be weighed, of course, along with other important factors such as cost, re-usability of materials, and teacher attitudes toward the programs. But assuming such factors are relatively equal, appraisal of results such as those included in Figure 3 should aid the evaluator.

Seven Steps

Recapitulating, then, here are the steps involved in conducting a program fair evaluation of two instructional programs:

1. Identify objectives (a) common to both programs, (b) unique to one program, and (c) unique to the other program.

2. Collect or construct test items based on the three sets of objectives.

3. Combine the test items into a three-part examination.

4. Assign estimates of importance, as explicitly as possible, to the three sets of objectives.

5. Administer each of the instructional programs to one of two randomly selected groups of appropriate learners.

6. Administer the three-part examination to both groups.

7. Appraise results and reach a decision regarding the preferred program.

The summative evaluation scheme described here for comparing programs with dissimilar objectives is relatively straightforward and can, therefore, be carried out with little difficulty. It can be applied to the evaluation of short-duration instructional sequences or to programs of much greater magnitude.[5] While conceptually simple, it provides

---

[5] One of the best examples of this approach has been conducted in Britain: W. A. Brownell "Two Recent Studies of Arithmetical Learning in British Schools", Address presented to the California Educational Research Association, March 16, 1968, Berkeley, California.

the evaluator with opportunities, if he wishes, to employ sophisticated quantification schemes to deal with such problems as value weightings of objectives. The major purpose of the procedure, of course, is to make evaluators attentive to the differential consequences of employing instructional programs which were designed with different intentions.